Identifying lineage effects when controlling for population structure improves power in bacterial association studies

Article (Accepted Version)

Published in final edited form as:
Nat Microbiol. ; 1: 16041. doi:10.1038/nmicrobiol.2016.41.

Identifying lineage effects when controlling for population structure improves power in bacterial association studies

Sarah G. Earle[#1], **Chieh-Hsi Wu**[#1], **Jane Charlesworth**[#1], **Nicole Stoesser**[1], **N. Claire Gordon**[1], **Timothy M. Walker**[1], **Chris C. A. Spencer**[2], **Zamin Iqbal**[2], **David A. Clifton**[3], **Katie L. Hopkins**[4], **Neil Woodford**[4], **E. Grace Smith**[5], **Nazir Ismail**[6,7], **Martin J. Llewelyn**[8], **Tim E. Peto**[1], **Derrick W. Crook**[1], **Gil McVean**[2,9], **A. Sarah Walker**[1], and **Daniel J. Wilson**[1,2,⋆]

[1]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK

[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

[3]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK

[4]Antimicrobial Resistance and Healthcare Associated Infections Reference Unit, Public Health England, London NW9 5EQ, UK

[5]Public Health England, West Midlands Public Health Laboratory, Heartlands Hospital, Birmingham B9 5SS, UK

[6]Centre for Tuberculosis, National Institute for Communicable Diseases, Johannesburg 2131 South Africa

[7]Department of Medical Microbiology, University of Pretoria, Pretoria, South Africa

[8]Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton BN2 5BE, UK

[9]The Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7FZ, UK

[#] These authors contributed equally to this work.

[⋆]Correspondence and requests for materials should be addressed to D.J.W. daniel.wilson@ndm.ox.ac.uk.

Accession codes. All genomes were deposited in NCBI and EBI short read archives under BioProject accession nos. PRJNA306133 (*E. coli* and *K. pneumoniae*), PRJNA308279 (*M. tuberculosis*) and PRJNA308283 (*S. aureus*). Individual BioSample accession numbers and antimicrobial resistance phenotypes are detailed in Supplementary Data 1.

Author contributions

S.G.E., C.-H.W., J.C. and D.J.W. designed the study, developed the methods, performed the analysis, interpreted the results and wrote the manuscript. Z.I. and D.A.C. assisted the analysis and commented on the manuscript. N.S., N.C.G., T.M.W., K.L.H., N.W., E.G.S., N.I., M.J.L., T.E.P. and D.W.C. designed and implemented isolate collection, drug susceptibility testing and whole-genome sequencing, and assisted with interpretation. C.C.A.S., G.M. and A.S.W. assisted with methods development and writing of the manuscript.

Additional information

Competing interests

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

Abstract

Bacteria pose unique challenges for genome-wide association studies because of strong structuring into distinct strains and substantial linkage disequilibrium across the genome1,2. Although methods developed for human studies can correct for strain structure3,4, this risks considerable loss-of-power because genetic differences between strains often contribute substantial phenotypic variability5. Here, we propose a new method that captures lineage-level associations even when locus-specific associations cannot be fine-mapped. We demonstrate its ability to detect genes and genetic variants underlying resistance to 17 antimicrobials in 3,144 isolates from four taxonomically diverse clonal and recombining bacteria: *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Escherichia coli* and *Klebsiella pneumoniae*. Strong selection, recombination and penetrance confer high power to recover known antimicrobial resistance mechanisms and reveal a candidate association between the outer membrane porin *nmpC* and cefazolin resistance in *E. coli*. Hence, our method pinpoints locus-specific effects where possible and boosts power by detecting lineage-level differences when fine-mapping is intractable.

Mapping genetic variants underlying bacterial phenotypic variability is of great interest owing to the fundamental role of bacteria ecologically, industrially and in the global burden of disease6–8. Hospital-associated infections including *Staphylococcus aureus*, *Escherichia coli* and *Klebsiella pneumoniae* represent a serious threat to the safe provision of healthcare9,10, while the *Mycobacterium tuberculosis* pandemic remains a major global health challenge11. Treatment options continue to be eroded by the spread of antimicrobial resistance, with some strains resistant even to antimicrobials of last resort12.

Genome-wide association studies (GWASs) offer new opportunities to map bacterial phenotypes through inexpensive sequencing of entire genomes, enabling direct analysis of causal loci and functional validation via well-developed molecular approaches2,13–22. However, bacterial populations typically exhibit genome-wide linkage disequilibrium and strong structuring into geographically widespread genetic lineages or strains that are probably maintained by natural selection1,5. Approaches to controlling for this population structure have allowed for systematic phenotypic differences based on cluster membership15,16 or, in clonal species, phylogenetic history13,19–21. However, these and other approaches common in human GWASs3,4 risk masking causal variants because differences between strains account for large proportions of both phenotypic and genetic variability.

Here, we describe a new approach for controlling bacterial population structure that boosts power by recovering signals of lineage-level associations when associations cannot be pinpointed to individual loci because of strong population structure, strong linkage disequilibrium and a lack of homoplasy. We base our approach on linear mixed models (LMMs), which can control for close relatedness within samples by capturing the fine structure of populations more faithfully than other approaches23 and enjoy greater applicability than phylogenetic methods because recombination is evident in most bacteria24,25. Our approach offers biological insights into strain-level differences and identifies groups of loci that are collectively significant, even when individually insignificant, without sacrificing the power to detect locus-specific associations.

Controlling for population structure aims to avoid spurious associations arising from (1) linkage disequilibrium with genuine causal variants that are population-stratified, (2) uncontrolled environmental variables that are population-stratified and (3) population-stratified differences in sampling3. In the four species we investigated, we observed genome-wide linkage disequilibrium and strong population structure, with the first ten principal components (PCs)26 explaining 70–93% of genetic variation, compared with 27% in human chromosome 1 (Supplementary Fig. 1). Controlling artefacts arising from population structure therefore risks a loss of power to detect genuine associations in this large proportion of population-stratified loci.

For example, we investigated associations between fusidic acid resistance and the presence or absence of short 31 bp haplotypes or 'kmers' in *S. aureus* (see Methods and Supplementary Fig. 2). The kmer approach aims to capture resistance encoded by substitutions in the core genome, the presence of mobile accessory genes, or both13. Kmers linked to the presence of *fusC*, a mobile element-associated resistance-conferring gene whose product prevents fusidic acid interacting with its target EF-G (ref. 27), showed the strongest genome-wide association by χ^2 test ($P = 10^{-122}$).

However, *fusC*-encoded resistance was observed exclusively within strains ST-1 and ST-8. Thus, controlling for population structure using LMM28 reduced the significance to $P = 10^{-39}$, below other loci (Fig. 1a and Supplementary Fig. 3). Kmers capturing resistance-conferring substitutions in *fusA*, which encodes EF-G, were propelled to greater significance, because these low-frequency variants were unstratified and LMM improves power in the presence of polygenic effects29 ($P = 10^{-11}$ by χ^2 test, $P = 10^{-157}$ by LMM). However, *fusA* variants explain only half as much resistance as *fusC* overall.

Although kmers linked to *fusC* did not suffer an outright loss of significance, as penetrance (proportion of *fusC* carriers expressing resistance) was very high, simulations show that for phenotypes with modest effect sizes (for example, odds ratios of 3), controlling for population structure risks loss of genome-wide significance at 59, 75, 99 and 99% of high-frequency causal variants in *M. tuberculosis* ($n = 1{,}573$), *S. aureus* ($n = 992$), *E. coli* ($n = 241$) and *K. pneumoniae* ($n = 176$) simulations, respectively, with the power loss being greatest when the sample size is low and the number of variants is high (Fig. 2a and Supplementary Fig. 4a).

Methods to limit loss of power such as 'leave-one-chromosome-out'29 are impractical in bacteria, which typically have one chromosome. Instead, we developed a method to recover information discarded when controlling for population structure. In cases where population stratification reduces the power to detect locus-specific associations, our method infers lineage-specific associations, similar to a phylogenetic regression30,31, without sacrificing the power to detect locus-specific associations when possible.

We observed that leading principal components tend to correspond to major lineages in bacterial genealogies (or 'clonal frames'32) despite substantial differences in recombination rates (Fig. 1b and Supplementary Fig. 5), reflecting an underlying relationship between genealogical history and principal component analysis33. Principal components are

commonly used to control for population structure by including leading principal components as fixed effects in a regression26. The regression coefficients estimated for principal components could therefore be interpreted as capturing lineage-level phenotypic differences, and each principal component tested for an effect on the phenotype. Because principal components are guaranteed to be uncorrelated, defining lineages in terms of principal components, rather than as phylogenetic branches or genetic clusters, minimizes the loss of power to detect lineage-level associations caused by correlations between lineages.

To identify lineage effects we exploited a connection between principal components and LMMs. In an LMM, every locus is included as a random effect in a regression. This is equivalent to including every principal component in the regression as a random effect34. We thus decomposed the random effects estimated by the LMM to obtain coefficients and standard errors for every principal component (see Methods). We then used a Wald test35 to assess the significance of the association between each lineage and the phenotype.

Our method, implemented in the R package *bugwas*, revealed strong signals of association between fusidic acid resistance and lineages including PC-6 and PC-9 ($P = 10^{-70}$), comparable in significance to the low-frequency variants at *fusA* (Fig. 1c and Supplementary Fig. 6). We next reassessed locus-specific effects by assigning variants to lineages according to the principal component to which they were most correlated, then comparing the significance of variants within lineages. This showed that *fusC* and variants in linkage disequilibrium with *fusC* accounted for the strongest signals within PC-6 and PC-9 ($P = 10^{-34}$ and 10^{-45} respectively, Fig. 1d), with the strongest locus-specific associations localized to a 20 kb region containing the staphylococcal cassette chromosome (SCC), the most significant hit mapping to the gene adjacent to *fusC*. Thus, identifying loci contributing to the most significant lineages provides an alternative to prioritizing variants for follow-up based solely on locus-specific significance.

In simulations, our method was able to recover signals of lineage-level associations in cases where significance at individual loci was lost by controlling for population structure, increasing the power 2.5-fold (*M. tuberculosis*) to 22.0-fold (*E. coli*) (Fig. 2a and Supplementary Fig. 4a). LMM reduced the number of falsely detected single nucleotide polymorphisms (SNPs) by 30-fold (*K. pneumoniae*) to 3,600-fold (*S. aureus*). However, fine-mapping of causal variants to specific chromosomal regions frequently suffered from genome-wide linkage disequilibrium, because linkage disequilibrium is not generally organized into physically linked blocks along the chromosome (Fig. 2b and Supplementary Fig. 4b), underlining the importance of recovering power by interpreting lineage effects.

We noted a trade-off to interpreting lineage effects, because they are susceptible to confounding with population-stratified differences in environment or sampling (Supplementary Fig. 7). Therefore, non-random associations between lineages and uncontrolled variables that influence phenotype risk false detection of lineage-level differences.

Confronted with a strong population structure and genome-wide linkage disequilibrium in bacteria, we wished to test empirically the ability of GWASs to pinpoint genuine causal variants more generally. We therefore conducted 26 GWASs for resistance to 17 antimicrobials in 3,144 isolates across the major pathogens *M. tuberculosis*36, *S. aureus*37, *E. coli* and *K. pneumoniae*38 (Supplementary Fig. 8).

We supplemented the kmer approach by surveying the variation in SNPs and gene presence or absence. We imputed missing SNP calls by reconstructing the clonal frame followed by ancestral state reconstruction, an approach that generally outperformed imputation using Beagle (Supplementary Table 1, see Methods).

Correlated phenotypes caused by the presence of multi-drug-resistant isolates led to significant results in unexpected loci or regions in some analyses. A combination of first-line drug regimens contributes to multi-drug resistance co-occurrence in *M. tuberculosis*, which led to spurious associations as the top hit before controlling for population structure between ethambutol and pyrazinamide resistance and SNPs in rifampicin resistance-conferring *rpoB*. Even after controlling for population structure, these associations remained genome-wide significant at $P = 10^{-45}$ and $P = 10^{-54}$.

Antimicrobial resistance has arisen over 20 times per drug in the *M. tuberculosis* tree, through frequent convergent evolution (Supplementary Fig. 4c and Supplementary Fig. 8). Within a single gene, such as *rpoB*, there are multiple targets for selection. Both SNP and kmer-based approaches correctly identified variants in known resistance-causing codons, but greater significance was attained in the latter because the targets for selection were typically within 31 bp (Supplementary Fig. 9a). In these cases, absence of the wild-type allele was found to confer resistance, with power gained by pooling over the alternative mutant alleles.

For each drug and species, we evaluated whether the most significant hit identified by GWAS matched a known causal variant36–38 (Supplementary Table 2). By this measure, the performance of GWASs across species was very good, identifying genuine causal loci or regions in physical linkage with those loci for antimicrobial resistance in 25/26 cases for the SNP and gene approach and the kmer approach after controlling for population structure (Table 1 and Supplementary Table 3). For accessory genes such as β-lactamases, in particular, mobile element-associated regions of linkage disequilibrium were often detected together with the causal locus (Supplementary Fig. 9b).

Genuine resistance-conferring variants were detected in all but one study, demonstrating that the high accuracy attained in predicting antimicrobial resistance phenotypes from genotypes known from the literature37,39 is mirrored by good power to map the genotypes that confer antimicrobial resistance phenotypes using GWASs. However, these results also reflect the extraordinary selection pressures exerted by antimicrobials. High homoplasy at resistance-conferring loci caused by repeat mutation and recombination breaks down linkage disequilibrium, assisting mapping (Fig. 2c and Supplementary Fig. 4c).

For one drug, cefazolin, in *E. coli*, we identified a variation in the presence of an unexpected gene as the most strongly associated with resistance, *nmpC* ($P = 10^{-12.4}$). This gene encodes an outer membrane porin over-represented in susceptible individuals. Permeability in the

Europe PMC Funders Author Manuscripts

Salmonella typhimurium homologue mediates resistance to other cephalosporin β-lactams40, making this a strong candidate for a novel resistance-conferring mechanism discovered in *E. coli*.

Population structure presents the greatest challenge for GWASs in bacteria, because of the inherent trade-off between the power to detect genuine associations of population-stratified variants and robustness to unmeasured, population-stratified confounders. By introducing a test for lineage-specific associations, we allow these signals to be recovered even in the absence of homoplasy, while acknowledging the increased risk of confounding. Detecting lineage effects is valuable, because characterizing phenotypic variability in terms of strain-level differences is helpful for biological understanding and it permits the prediction of traits, including clinically actionable phenotypes, from strain designation.

Identifying loci that contribute to the most significant lineage-level associations offers flexibility in the interpretation of bacterial GWASs, where it will often be difficult to pinpoint significance to individual locus effects and where linkage disequilibrium can make the fine-mapping of causal loci a genome-wide problem. Loci can be prioritized for follow-up by identifying groups of lineage-associated variants that collectively show a strong signal of phenotypic association, but which cannot be distinguished statistically. This strategy provides an alternative to prioritizing variants based solely on locus-specific significance, but it carries risks, because lineage-associated effects are more susceptible to confounding with population-stratified differences in environment or sampling. This trade-off between power and robustness underlines the importance of functional validation for bacterial GWASs going forward.

Methods

Linear mixed model

In the LMM41–45, the phenotype is modelled as depending on the fixed effects of covariates including an intercept, the ‘foreground’ fixed effect of the locus whose individual contribution is to be tested, the ‘background’ random effects of all the loci whose cumulative contribution to phenotypic variability we will decompose into lineage-level effects, and the random effect of the environment:

$$\text{phenotype} = \text{covariates} + \text{foreground locus} + \text{background loci} + \text{environment}$$

Formally,

$$y_i = w_{i1}\alpha_1 + \ldots + w_{ic}\alpha_c + X_{il}\beta_l + X_{il}\gamma_1 + \ldots + X_{il}\gamma_L + \varepsilon_i$$

where there are n individuals, c covariates, L loci, l is the foreground locus, y_i is the phenotype in individual i, W_{ij} is covariate j in individual i, α_j is the effect of covariate j, X_{ij} is the genotype of locus j in individual i, β_l is the foreground effect of locus l, γ_j is the background effect of locus j and ε_i is the effect of the environment (or error) on individual i. Biallelic genotypes are numerically encoded as $-f_j$ (common allele) or $1-f_j$ (rare allele),

where f_j is the frequency of the rare allele at locus j. This convention ensures that the mean value of X_{ij} over individuals i is zero for any locus j. Because triallelic and tetrallelic loci are rare, we use only biallelic loci to model background effects. When the foreground locus is triallelic ($K = 3$) or tetrallelic ($K = 4$), the genotype in individual i is encoded as a vector indicating the presence (1) or absence (0) of the first ($K - 1$) alleles and β_l becomes a vector of length ($K - 1$).

Treating the background effects of the loci as random effects means the precise values of coefficients γ_j are averaged. The γ_j are assumed to follow independent normal distributions with common mean 0 and variance $\lambda\tau^{-1}$ to be estimated. As most loci are expected to have little or no effect on a particular phenotype, this tends to constrain the magnitude of the background effect sizes to be small. The environmental effects are also treated as random effects assumed to follow independent normal distributions with mean 0 and variance τ^{-1}. The model can be rewritten in matrix form as

$$\mathbf{y} = \mathbf{W}\alpha + \mathbf{X}_{\cdot l}\beta_l + \mathbf{u} + \varepsilon$$

with

$$\mathbf{u} = \mathbf{X}_{\cdot 1}\gamma_1 + \ldots + \mathbf{X}_{\cdot L}\gamma_L$$

$$\boldsymbol{\mu} \sim \mathrm{MVN}_n\left(0, \lambda\tau^{-1}\mathbf{K}\right)$$

$$\varepsilon \sim \mathrm{MVN}_n\left(0, \tau^{-1}\mathbf{I}_n\right)$$

where **u** represents the cumulative background effects of the loci, MVN denotes the multivariate normal distribution, $\mathbf{I}_n$ is an $n \times n$ identity matrix, and **K** is an $n \times n$ relatedness matrix defined as $\mathbf{K} = \mathbf{X}\mathbf{X}'$, which captures the genetic covariance between individuals.

Testing for locus effects

To assess the significance of the effect of an individual locus l on the phenotype, controlling for population structure and background genetic effects, the parameters of the linear mixed model $\alpha_1 \ldots \alpha_c$, β_l, λ and τ were estimated by maximum likelihood, and a likelihood ratio test with ($K - 1$) degrees of freedom was performed against the null hypothesis that $\beta_l = 0$ using the software GEMMA28.

Testing for lineage effects

Because controlling for population structure drastically reduces the power at population-stratified variants, and because a large proportion of variants are typically population-stratified in bacteria, we recovered information from the LMM regarding lineage-level differences in phenotype.

We defined lineages using principal components because we observed that principal components tend to trace paths through the clonal frame genealogy corresponding to recognizable lineages (as seen by the branch colouring in Fig. 1b and Supplementary Fig. 5) and because principal components are mutually uncorrelated, minimizing loss of power to detect differences between lineages due to correlations. Principal components were computed based on biallelic SNPs using the R function prcomp(), producing an L by n loading matrix $\mathbf{D}$ and an n by n score matrix $\mathbf{T}$ where $\mathbf{T} = \mathbf{X}\,\mathbf{D}$. D_{ij} records the contribution of biallelic SNP i to the definition of principal component j, while T_{ij} represents the projection of individual i onto principal component j.

Point estimates and standard errors for the background locus effects are usually overlooked because the assumed normal distribution with common mean 0 and variance $\lambda\tau^{-1}$ tends to cause them to be small in magnitude and not significantly different from zero. However, cumulatively, the background locus effects can capture systematic phenotypic differences between lineages. We therefore recovered the post-data distribution (equivalent to an empirical Bayes posterior distribution) of the background locus random effects, $\boldsymbol{\gamma}$, from the LMM, and reinterpreted it in terms of lineage-level differences in phenotype.

Empirically, we found that the post-data distribution of the background random effects was generally insensitive to the identity of the foreground locus and comparable under the null hypothesis ($\beta_l = 0$). We therefore calculated the mean and variance–covariance matrix of the multivariate normal post-data distribution of $\boldsymbol{\gamma}$ in the LMM null model. These are equivalent to those of a ridge regression46 and were computed as

$$\mu=(\mathbf{X}'\mathbf{X}+1/\lambda\mathbf{I}_L)^{-1}\mathbf{X}'\mathrm{y}\ \text{and}\ \sum(\tau\mathbf{X}'\mathbf{X}+1/\lambda\mathbf{I}_L)^{-1}$$

respectively. Both λ and τ were estimated by GEMMA under the LMM null model. Using the inverse transformation of the biallelic variants from PCA, $\mathbf{X} = \mathbf{T}\mathbf{D}^{-1}$, the background random effects can be rewritten in terms of the contribution of the n principal components:

$$\begin{aligned}\mathbf{u}&=\mathbf{X}_{\cdot 1}\gamma_1+\ldots+\mathbf{X}_{\cdot L}\gamma_L\\&=\mathbf{X}\boldsymbol{\gamma}=\mathbf{T}\mathbf{D}^{-1}\boldsymbol{\gamma}=\mathbf{T}\mathbf{g}\\&=\mathbf{T}_{\cdot 1}g_1+\ldots+\mathbf{T}_{\cdot n}g_n\end{aligned}$$

where $\mathbf{g} = \mathbf{D}^{-1}\boldsymbol{\gamma}$, and g_j is the background effect of principal component j on the phenotype. We computed the mean and variance of the post-data distribution of $\mathbf{g}$ as $\mathbf{m} = \mathbf{D}^{-1}\boldsymbol{\mu}$ and $\mathbf{S} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}$, respectively, using the affine transformation for a multivariate normal distribution. To test the null hypothesis of no background effect of principal component j (that is, $g_j = 0$), we used a Wald test with test statistic $w_j=m_j^2/S_{jj}$, which we compared against a χ^2 distribution with one degree of freedom to obtain a P value.

Although we identified and tested for lineage effects in the LMM setting, lineage effects could also be identified and tested for by interpreting the coefficients of leading principal components or genetic cluster membership included as fixed effects in a regression, both of which represent alternative methods for controlling for population structure.

Identifying non genome-wide principal components

Some principal components capture variation localized to particular areas of the genome. We identified non genome-wide principal components by testing for spatial heterogeneity of the loading matrix **W** for biallelic SNPs across the genome. SNPs were grouped into 20 contiguous bins (indexed by j) of nearly equal sizes N_j, and the mean O_{ij} and variance V_{ij} in the absolute value of the SNP loadings for principal component i in bin j were calculated, as well as the mean absolute value E_i of the SNP loadings for principal component i across all SNPs. The null hypothesis of no heterogeneity was assessed by comparing the test statistic $\chi_i^2 = \Sigma_j (O_{ij} - E_i)^2/(V_{ij}/N_j)$ to a χ^2 distribution with degrees of freedom equal to the number of bins minus one to obtain a P value.

Antimicrobial resistance testing, genome sequencing and SNP calling

We investigated 241 *E. coli* and 176 *K. pneumoniae* UK clinical isolates newly reported here, together with 992 *S. aureus* and 1,735 *M. tuberculosis* isolates reported previously36,37. All isolates were tested for resistance to multiple antimicrobials based on routine clinical laboratory protocols, and DNA was extracted and sequenced on Illumina platforms as previously described36–38. We called SNPs using standard methods47,48, employing Stampy49 to map reads to reference strains CFT073 (genbank accession no. AE014075.1), MGH 78578 (CP000647.1), H37Rv (NC_000962.2) and MRSA252 (BX571856.1) for *E. coli*, *K. pneumoniae*, *M. tuberculosis* and *S. aureus*, respectively. The distributions of biallelic SNP frequencies are provided in Supplementary Table 4.

Defining the pan-genome

To investigate gene presence or absence we created a pan-genome for each set of isolates. To obtain whole genome assemblies, reads were *de novo* assembled using Velvet50. We annotated open reading frames on the *de novo* assemblies for each isolate. We then used the Bayesian gene-finding program Prodigal51 to identify a set of protein sequences for each *de novo* assembly. These annotated protein sequences were clustered using CD-hit52, with a clustering threshold of 70% identity across 70% of the longer sequence. We converted the output of CD-hit into a matrix of binary genotypes denoting the presence or absence of each gene cluster in each genome (Supplementary Fig. 2).

Kmer counting

Some diversity such as indels and repeats is difficult to capture using standard variant calling tools. To capture non-SNP variation, we pursued a kmer or word-based approach13 in which all unique 31 base haplotypes were counted from the sequencing reads using dsk53 following adaptor trimming and removal of duplicates and low-quality reads using Trimmomatic54. If a kmer was counted five or more times in an isolate, then it was counted as present; if not, it was treated as absent (Supplementary Fig. 2). This produced a deduplicated set of variably present kmers across the data set, with the presence or absence of each determined per isolate. The total number of SNPs, kmers and gene clusters per species can be found in Supplementary Table 5.

Phylogenetic inference

Maximum likelihood phylogenies were estimated for visualization and SNP imputation purposes using RAxML version 7.7.6 (ref. 55), with a general time reversible (GTR) model and no rate heterogeneity, using alignments from the mapped data based on biallelic sites, with non-biallelic sites being set to the reference.

SNP imputation

Because Illumina sequencing is inherently more error-prone than Sanger sequencing, strict filtering is required for reliable mapping-based SNP calling, contributing to a small but appreciable frequency of uncalled bases in the genome due to ambiguity or deletion. Restricting analysis to sites called in all genomes is undesirable, while ignoring uncalled sites by removing individuals with missing data at individual sites generates *P* values that cannot be validly compared between sites because they are calculated using data from differing sets of isolates.

SNP imputation is therefore generally considered necessary for GWASs56. We imputed missing base calls using two approaches, ClonalFrameML57 and Beagle56. Imputation using ClonalFrameML57 involves estimating the clonal frame by maximum likelihood58, then jointly reconstructing ancestral states and missing base calls by maximum likelihood utilizing the phylogeny reconstructed earlier59. To use Beagle, the mapped data were coded as haploid (one column per individual) and input as phased data56,60.

Testing imputation accuracy

To simulate data for testing imputation accuracy, 100 sequences were randomly sampled from each GWAS data set across the phylogeny. Maximum likelihood phylogenies were estimated for the 100 sequences of each species using RAxML55, as above. Any columns in the alignment corresponding to ambiguous bases in the reference genome were excluded. One round of imputation was performed using ClonalFrameML to produce complete data sets with no ambiguous bases (Ns), which were then treated as the truth for the purpose of testing. The empirical distributions of Ns per site in the data sets of 100 sequences were determined, and these were sampled with replacement to reintroduce Ns to the variable sites in 100 simulated data sets. These sequences were then imputed again using ClonalFrameML and Beagle. Accuracy was summarized per site as a function of the frequency of Ns per site and the minor allele frequency. Overall, ClonalFrameML was more accurate than Beagle, so ClonalFrameML was used for all GWAS analyses (Supplementary Table 1).

Calculating association statistics before controlling population structure

We wished to compare the significance of associations before and after controlling for population structure. For the SNP and gene presence or absence data, an association between each SNP or gene and the phenotype was tested by logistic regression implemented in R. For the kmer analyses, an association between the presence or absence of each kmer was tested using a χ^2 test implemented in C++. For each variant a *P* value was computed.

Correction for multiple testing

Multiple testing was accounted for by applying a Bonferroni correction61; the individual locus effect of a variant (SNP, gene or kmer) was considered significant if its P value was smaller than α/n_p, where we took $\alpha = 0.05$ to be the genome-wide false-positive rate and n_p to be the number of SNPs and genes, or kmers, with unique phylogenetic patterns, that is, unique partitions of individuals according to allele membership. Because the phenotypic contribution of multiple variants with identical phylogenetic patterns cannot be disentangled statistically, we found that pooling such variants improved the power by demanding a less conservative Bonferroni correction than correcting for the total number of variants (Supplementary Fig. 10).

The genome-wide $-\log_{10} P$ value threshold for SNPs and genes (or kmers) was 6.1 (7.3) for *S. aureus* ciprofloxacin, erythromycin, fusidic acid, gentamicin, penicillin, methicillin, tetracycline and rifampicin, 5.9 (6.7) for *S. aureus* trimethoprim, 6.5 (7.3) for all antimicrobials tested for *E. coli*, 6.6 (7.3) for all antimicrobials tested for *K. pneumoniae* and 5.0 (7.6) for all antimicrobials tested for *M. tuberculosis*. We also accounted for multiple testing of lineage effects by applying a Bonferroni correction for the number of principal components, which equals the sample size n.

Running GEMMA

For the analyses of SNPs, genes and kmers, we computed the relatedness matrix **K** from biallelic SNPs only. We tested for foreground effects at all biallelic, triallelic and tetrallelic SNPs, genes and kmers. GEMMA was run using a minor allele frequency of 0 to include all SNPs. GEMMA was modified to output the ML log-likelihood under the null, and alternative and $-\log_{10} P$ values were calculated using R.

To perform LMM on tri- and tetra-allelic SNPs, each SNP was encoded as $K - 1$ binary columns corresponding to the first $K - 1$ alleles. For each column, an individual was encoded 1 if it contained that allele and 0 otherwise. The first column was input as the genotype, and the others as covariates into GEMMA. The log-likelihood of the null from the biallelic SNPs, together with the log-likelihood under the alternative for each of the SNPs, was used to calculate the P value per SNP.

Due to the large number of kmers present within each data set, it was not feasible to run LMM on all kmers. We therefore applied the LMM to the top 200,000 most significant kmers from the logistic regression, plus 200,000 randomly selected kmers of those remaining. The randomly selected kmers were used to indicate whether some were becoming relatively more significant than the top 200,000, providing a warning in the case where large numbers of kmers became significant only after controlling for population structure.

Variant annotation

SNPs were annotated in R using the reference fasta and genbank files to determine SNP type (synonymous, non-synonymous, nonsense, read-through and intergenic), the codon and codon position, reference and non-reference amino acid, gene name and gene product.

Europe PMC Funders Author Manuscripts

Unlike the SNP approach, where we can easily refer to the reference genome to find what gene the SNP is in and the effect that it may have, annotation of the kmers is more difficult. We used BLAST62 to identify the kmers in databases of annotated sequences. Each kmer was first annotated against a BLAST database created of all refseq genomes of the relevant genus on NCBI. This enabled automatic annotation of all kmers that gave a sufficiently small e-value against the genus-specific database. All kmers were also searched against the whole nucleotide NCBI database, first to compare and confirm the matches made against the first database and second to annotate the kmers that did not match anything in the within-genus database. Finally, when the resistance-determining mechanism was a SNP, the top 10,000 kmers were mapped to a relevant reference genome using Bowtie2 (ref. 63). This was used to determine whether the most significant kmers covered the position of the resistance-causing SNP or whether they were found elsewhere in the gene.

Genes were annotated for each CD-hit gene cluster by performing BLAST62 searches of each cluster sequence against a database of curated protein sequences downloaded from UNIPROT64.

Testing power by simulating phenotypes

To assess the performance of the method for controlling population structure, we performed 100 simulations per species. In each simulation, a biallelic SNP was chosen randomly (from those SNPs with minor allele frequency above 20%) to be the causal SNP. Binary phenotypes (case or control) were then simulated for each genome with case probabilities of 0.25 and 0.5, respectively, in individuals with the common and rare allele at the causal SNP (an odds ratio of 3). For each simulated data set, we tested for locus effects at every biallelic SNP, and for lineage effects at every principal component, as described above. The power to detect locus effects was defined as the proportion of simulations in which the causal SNP was found to have a significant locus effect. This was compared to a theoretically optimum power computed as the proportion of simulations in which the causal SNP was found to have a significant locus effect when population structure and multiple testing were not controlled for. The power to detect lineage effects was computed as the proportion of simulations in which the principal component most strongly correlated to the causal SNP was found to have a significant lineage effect. We defined fine mapping precision as the distance spanned by SNPs within two log-likelihoods of the most significant SNP in the test for locus effects, in those simulations in which the causal locus was genome-wide significant. We calculated the number of homoplasies per SNP by counting the number of branches in the phylogeny affected by a substitution based on the ClonalFrameML ancestral state reconstruction, and subtracting the minimum number of substitutions ($K - 1$).

Code availability

We have created an R package, *bugwas*, implementing our method for controlling population structure, and an end-to-end GWAS pipeline using R, Python and C++. Both can be downloaded from www.danielwilson.me.uk/virulogenomics.html.

Europe PMC Funders Author Manuscripts

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank J.-B. Veyrieras, D. Charlesworth and B. Charlesworth for comments on the manuscript, X. Zhou and M. Stephens for helping adapt their software, S. Niemann for assisting with tuberculosis isolates and X. Didelot, D. Falush, R. Bowden, S. Myers, J. Marchini, J. Pickrell, P. Visscher, A. Price and P. Donnelly for discussions. This study was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant and the UKCRC Modernising Medical Microbiology Consortium, the latter funded under the UKCRC Translational Infection Research Initiative supported by the Medical Research Council, the Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the UK Department of Health (grant no. G0800778) and the Wellcome Trust (grant no. 087646/Z/08/Z). T.M.W. is an MRC research training fellow. C.C.A.S. was supported by a Wellcome Trust Career Development Fellowship (grant no. 097364/Z/11/Z). D.A.C. is funded by the Royal Academy of Engineering and an EPSRC Healthcare Technologies Challenge Award. T.E.P. and D.W.C. are NIHR Senior Investigators. G.M. is supported by a Wellcome Trust Investigator Award (grant no. 100956/Z/13/Z). D.J.W. and Z.I. are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and the Royal Society (grants nos. 101237/Z/13/Z and 102541/Z/13/Z).

References

1. Feil EJ, Spratt BG. Recombination and the structures of bacterial pathogens. Annu Rev Microbiol. 2001; 55:561–590. [PubMed: 11544367]
2. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006; 14:353–355. [PubMed: 16782339]
3. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Rev Genet. 2009; 10:681–690. [PubMed: 19763151]
4. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]
5. Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Rev Microbiol. 2014; 12:263–273. [PubMed: 24590245]
6. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA. 1998; 95:6578–6583. [PubMed: 9618454]
7. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008; 320:1034–1039. [PubMed: 18497287]
8. World Health Organization. The Global Burden of Disease: 2004 Update. 2008. http://www.who.int/healthinfo/global_burden_disease
9. Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 2010; 74:417–433. [PubMed: 20805405]
10. European Centre for Disease Prevention and Control. Surveillance of Surgical-Site Infections in Europe, 2008–2009. 2012. http://www.ecdc.europa.eu/en/publications/Publications/120215_SUR_SSI_2008-2009.pdf
11. World Health Organization. Global Tuberculosis Report 2014. 2014. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf
12. World Health Organization. Antimicrobial Resistance: A Global Report on Surveillance. 2014. http://www.who.int/iris/bitstream/10665/112642/1/9789241564748_eng.pdf
13. Sheppard SK, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci USA. 2013; 110:11923–11927. [PubMed: 23818615]
14. Alam MT, et al. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. Genome Biol Evol. 2014; 6:1174–1185. [PubMed: 24787619]
15. Laabei M, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014; 24:839–849. [PubMed: 24717264]

16. Chewapreecha C, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014; 10:e1004547. [PubMed: 25101644]
17. Salipante SJ, et al. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. Genome Res. 2014; 25:119–128. [PubMed: 25373147]
18. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014; 6:109. [PubMed: 25593593]
19. Fahrat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Med. 2014; 6:101. [PubMed: 25484920]
20. Hall BG. SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignments. PLoS ONE. 2014; 9:e90490. [PubMed: 24587377]
21. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015; 25:17–24. [PubMed: 25835153]
22. Holt KE, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci USA. 2015; 112:E3574–E3581. [PubMed: 26100894]
23. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nature Rev Genet. 2010; 11:459–463. [PubMed: 20548291]
24. Perez-Losada M, et al. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol. 2006; 6:97–112. [PubMed: 16503511]
25. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archeae. IMSE J. 2009; 3:199–208.
26. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 2006; 38:904–909. [PubMed: 16862161]
27. O'Neill AJ, McLaws F, Kahlmeter G, Henriksen AS, Chopra I. Genetic basis of resistance to fusidic acid in staphylococci. Antimicrob Agents Chemother. 2007; 51:1737–1740. [PubMed: 17325218]
28. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genet. 2012; 44:821–824. [PubMed: 22706312]
29. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature Genet. 2014; 46:100–106. [PubMed: 24473328]
30. Grafen A. The phylogenetic regression. Phil Trans R Soc Lond B. 1989; 326:119–157. [PubMed: 2575770]
31. Martins EP, Hansen TF. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat. 1997; 149:646–667.
32. Milkman R, Bridges MM. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics. 1990; 126:505–517. [PubMed: 1979037]
33. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009; 5:e1000686. [PubMed: 19834557]
34. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Stat Sci. 2009; 24:451–471.
35. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc. 1943; 54:426–482.
36. Walker TM, et al. Whole genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015; 15:1193–1202. [PubMed: 26116186]
37. Gordon NC, et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. J Clin Microbiol. 2014; 52:1182–1191. [PubMed: 24501024]

38. Stoesser N, et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genome sequence data. J Antimicrob Chemother. 2013; 68:2234–2244. [PubMed: 23722448]
39. Bradley P, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis* . Nature Commun. 2015; 6:10063. [PubMed: 26686880]
40. Sun S, Berg OG, Roth JR, Andersson DI. Contribution of gene amplification to evolution of increased antibiotic resistance in *Salmonella typhimurium* . Genetics. 2009; 182:1183–1195. [PubMed: 19474201]
41. Yu J, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genet. 2006; 38:203–208. [PubMed: 16380716]
42. Kang HM, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]
43. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genet. 2010; 42:348–354. [PubMed: 20208533]
44. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nature Methods. 2011; 8:833–835. [PubMed: 21892150]
45. Listgarten J, et al. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012; 9:525–526. [PubMed: 22669648]
46. O'Hagan, A.; Forster, J. Kendall 's Advanced Theory of Statistics Volume 2B Bayesian Inference. 2nd edn. Vol. Ch 11. Wiley-Blackwell; 2010.
47. Eyre DW, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium diffcile* for outbreak detection and surveillance. BMJ Open. 2012; 2:e001124.
48. Everitt RG, et al. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus* . Nature Commun. 2014; 5:3956. [PubMed: 24853639]
49. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011; 21:936–939. [PubMed: 20980556]
50. Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
51. Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11:119. [PubMed: 20211023]
52. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22:1658–1659. [PubMed: 16731699]
53. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. Bioinformatics. 2013; 29:652–653. [PubMed: 23325618]
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114–2120. [PubMed: 24695404]
55. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. [PubMed: 24451623]
56. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81:1084–1097. [PubMed: 17924348]
57. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015; 11:e1004041. [PubMed: 25675341]
58. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. mBio. 2014; 5:e02158–14. [PubMed: 25425237]
59. Pupko T, Pe'er I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 2000; 17:890–896. [PubMed: 10833195]
60. Yahara K, Didelot X, Ansari M, Sheppard SK, Falush D. Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol. 2014; 31:1593–1605. [PubMed: 24586045]
61. Dunn OJ. Estimation of the medians for dependent variables. Ann Math Stat. 1959; 30:192–197.
62. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:431. [PubMed: 20021653]

63. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357–359. [PubMed: 22388286]

64. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–D212. [PubMed: 25348405]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

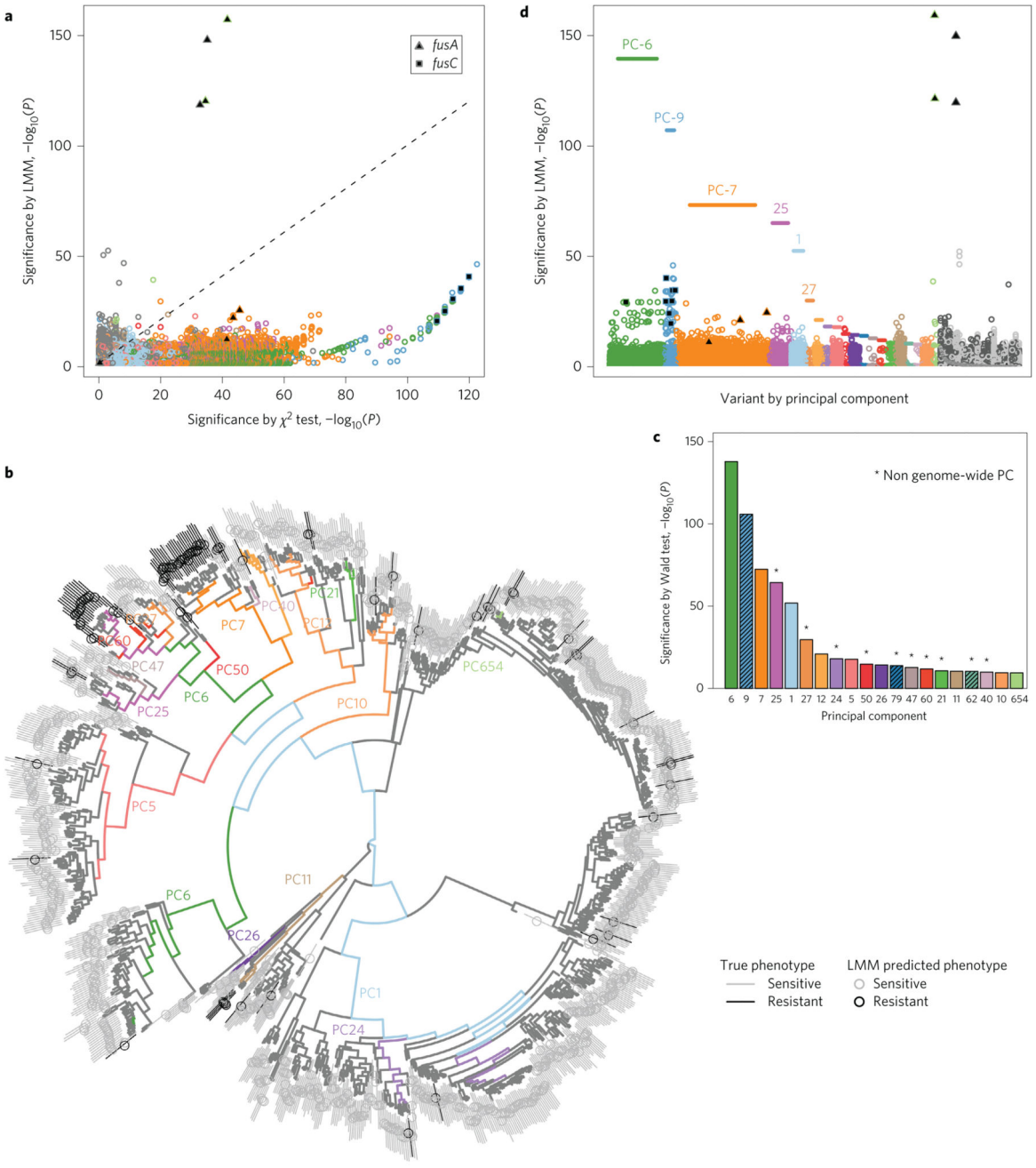

Figure 1. Controlling for population structure in bacterial GWASs for fusidic acid resistance in *S. aureus*.

a, Effect of controlling for population structure using LMM on the significance of the presence or absence of 31 bp kmers. The 200,000 most-significant kmers prior to control for population structure and a random 200,000 are plotted. Each kmer is colour-coded according to the principal component to which it is most strongly correlated, and grey if it is not most strongly correlated to one of the 20 most significant principal components. **b**, Principal components correspond to lineages in the clonal genealogy. Branches are colour-coded by one of the 20 most significant principal components to which they are most correlated.

Individual genomes are colour-coded with black or grey lines to indicate fusidic acid resistance and susceptibility, respectively. The circle passing through the line is colour-coded to indicate the phenotype predicted by the LMM. **c**, Wald tests of significance of lineage-specific associations. Some principal components, for example, PC-9, are hashed to indicate that no branch in the clonal genealogy was most strongly correlated with it. Asterisks above the bars, for example PC-25, indicate evidence for lineages associated with particular genomic regions. **d**, Manhattan plot showing significance of unique variants after controlling for population structure, with variants clustered by principal component. The horizontal ordering is randomized. This allows identification of the variants corresponding to the most significant lineage-specific associations.

Europe PMC Funders Author Manuscripts

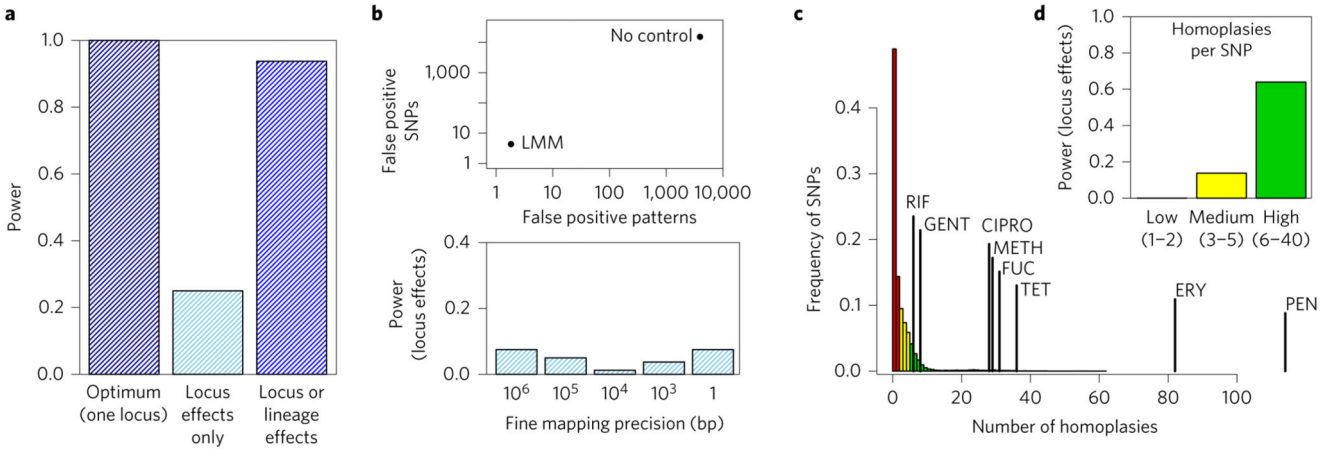

Figure 2. Power, false positives, fine mapping and homoplasy in *S. aureus*. Simulation results.

a, Controlling for population structure and multiple testing lead to a drastic reduction in power to detect locus effects, compared with the theoretical optimum power for a single locus. The Wald test improves the power several-fold by detecting lineage-specific effects. **b**, Top: mean numbers of false-positive SNPs and patterns (that is, unique distributions of SNP alleles among individuals) are drastically reduced by controlling population structure with LMM. Bottom: fine mapping precision is very coarse owing to genome-wide linkage disequilibrium. Interpreting lineage effects is useful when the locus-specific signal cannot be fine-mapped. **c**, Number of times that common SNPs (minor allele frequency (MAF) > 20%) and antibiotic resistance phenotypes have emerged on the phylogeny. **d**, When homoplasy is high, the power to detect locus effects is much improved, explaining the good power to map antibiotic resistance phenotypes. In the simulations, causal loci were selected at random from high-frequency SNPs (MAF > 20%) in the $n = 992$ isolates and phenotypes simulated per genome with case probabilities of 0.25 and 0.5 for the common and rare alleles, respectively (odds ratio of 3). Genome-wide significance (to detect locus effects) was based on a Bonferroni-corrected P value threshold of α, equal to 0.05 divided by the number of SNP patterns.

Europe PMC Funders Author Manuscripts

Nat Microbiol. Author manuscript; available in PMC 2016 October 04.

Table 1

Number of resistant and sensitive isolates by species and antibiotics, known mechanisms of resistance and main results.

Antibiotic	R	S	Resistance mechanism	Resistance determined by	SNP/gene rank	SNP/gene LMM rank	Kmer rank	Kmer LMM rank
E. coli								
Ampicillin	189	52	β-lactamase genes bla_{TEM}	Gene presence	1	1	6 (tnp) *	6 (tnp) *
Cefazolin	139	102	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	2 (*nmpC*) **	3 (*nmpC*) **	121,710 (nmpC) **	3,690 (nmpC) **
Cefuroxime	81	160	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	1	1	1,598 (162-192 upstream $bla_{CMY\text{-}2}$) *	470 (162-192 upstream $bla_{CMY\text{-}2}$) *
Ceftriaxone	55	186	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	1	1	1,403 (tnp) *	470 (tnp) *
Ciprofloxacin	91	150	SNPs in *gyrA*#, *gyrB*, *parC*## or *parE* or presence of PMQR‡	Gene presence or SNPs, or both	1##	1##	1##	1#
Gentamicin	48	193	*aac* (*aac(3)-II*), *ant*, *aph* or rRNA methylase	Gene presence	1	1	1	1
Tobramycin	67	174	*aac* (*aac(3)-II*), *ant* or rRNA methylase	Gene presence	1	1	1	1
K. pneumonia								
Cefazolin	53	123	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	1 + HP + *wbuC*	1	762 (tnp) *	837 (tnp) *
Cefuroxime	46	130	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	1 + HP + *wbuC*	1+ HP + *wbuC*	762 (tnp) *	1,480 (tnp) *
Ceftriaxone	35	141	β-lactamase genes $bla_{CTX\text{-}M}$	Gene presence	1 + HP + *wbuC*	1 + HP + *wbuC*	771 (tnp) *	812 (tnp) *
Ciprofloxacin	34	142	SNPs in *gyrA*, *gyrB*, *parC* or *parE* or presence of PMQR (*qnr-B1*#, *qnr-B19*##)	Gene presence or SNPs, or both	2# (tnp) *	2# (tnp) *	1,853## (tnp) *	4,427## (tnp) *
Gentamicin	31	145	*aac* (*acc(3)-II*), *ant*, *aph* or rRNA methylase	Gene presence	1	1	1	79 (*tmrB_2*) *
Tobramycin	36	140	*aac* (*acc(3)-II*), *ant* or rRNA methylase	Gene presence	1	1	1	1
M. tuberculosis								
Ethambutol	41	1,589	*embB*	SNPs	2 (*rpoB*) **	1	1	1
Isoniazid	239	1,470	*katG*, *fabG1*	SNPs	1	1	1	1
Pyrazinamide	45	1,662	*pncA*	SNPs	142 (*rpoB*) **	1	126 (*rpoB*) **	1
Rifampicin	86	1,487	*rpoB*	SNPs	1	1	1	1
S. aureus								
Ciprofloxacin	242	750	*grlA* or *gyrA*	SNPs	1	1	1	1

Europe PMC Funders Author Manuscripts

Antibiotic	R	S	Resistance mechanism	Resistance determined by	SNP/gene rank	SNP/gene LMM rank	Kmer rank	Kmer LMM rank
Erythromycin	216	776	*ermA*, *ermC*, *ermT* or *msrA*	Gene presence	1	1	1	1
Fusidic acid	84	908	SNPs in *fusA*[#] or presence of *fusB* or *fusC*[##]	Gene presence or SNPs, or both	4[##] (*SAS0037*)[*]	1[#]	75[##] (*SAS0040*)[*]	1[#]
Gentamicin	11	981	*aacA/aphD*	Gene presence	1 + GNAT acetyltransferase	1 +GNAT acetyltransferase	1 + 415 bases upstream to 100 bases downstream	1 + 415 bases upstream to 100 bases downstream
Penicillin	824	168	*blaZ*	Gene presence	1	1	2 (*blaI*)[*]	2 (*blaI*)[*]
Methicillin	216	776	*mecA*	Gene presence	1	1 + *mecR1*	1 + SCC*mec* genes	1 + SCC*mec* genes
Tetracycline	46	946	*tetK*, *tetL* or *tetM*	Gene presence	2 (*repC*)[*]	2 (*repC*)[*]	1 + plasmid genes	1 + plasmid genes
Trimethoprim	15	308	SNPs in *dfrB*, presence of *dfrG* or *dfrA*	Gene presence or SNPs, or both	1	1	1	1
Rifampicin	8	984	*rpoB*	SNPs	1	1	1	1

For each antibiotic, the most significant variant was the expected mechanism, unless indicated by *(most significant variant was in physical linkage (PL) with the expected mechanism) or **(most significant variant was not the expected mechanism or in PL with the expected mechanism). The rank of the most significant result for an expected causal mechanism for each GWAS is reported, plus, in brackets, the gene that was most significant when it was not causal. Where more than one gene or mechanism causes resistance, the variant we found is underlined, or referred to by [#] and [##]. R, resistant; S, sensitive; HP, hypothetical protein; tnp, transposase; PMQR, plasmid mediated quinoline resistance. See Supplementary Tables 3–6 for more detail.